This dissertation has been
microfilmed exactly as received

66-1839

SHINE II, Lester Caldwell, 1933-
THE RELATIVE EFFICIENCY OF TEST SELECTION
METHODS IN CROSS-VALIDATION ON GENERATED
DATA.

The Ohio State University, Ph.D., 1965
Psychology, experimental

University Microfilms, Inc., Ann Arbor, Michigan

THE RELATIVE EFFICIENCY OF TEST SELECTION METHODS
IN CROSS-VALIDATION ON GENERATED DATA

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the
Degree Doctor of Philosophy in the Graduate School of
The Ohio State University

By

LESTER CALDWELL SHINE II, B.B.A., M.A.

* * * * *

The Ohio State University
1965

Approved by

Adviser
Department of Psychology

PREFACE

This dissertation has grown out of the State Farm Research Project, The Ohio State University. Its purpose is to investigate what test selection methods work best under what conditions, and to thereby shed light upon one of the crucial problems in the State Farm Project. The results of this investigation are, of course, not limited to answering the questions of the Project alone.

The author wishes to gratefully acknowledge Drs. R. Wherry, C. Shartle, and J. Naylor, whose tremendous pool of knowledge made much of this dissertation possible.

CONTENTS

	Page
PREFACE	ii
ILLUSTRATIONS	iv
Chapter	
I. INTRODUCTION	1
II. METHOD	5
III. RESULTS	9
IV. DISCUSSION AND CONCLUSIONS	16
V. SUMMARY	20
REFERENCES	23
APPENDIX	25
AUTOBIOGRAPHY	33

ILLUSTRATIONS

Table	Page
1. The Four Factor Structures (<u>A</u> Matrices)	7
2. Cross-validation Data after Fisher Transformation . . .	11
3. ANOVA Source Table	13
4. Ordered Means from NxSxM Effect Table	14

Figure

1. NxSxM Interaction	15
--------------------------------	----

CHAPTER I

INTRODUCTION

The basic problem in test selection is one of selecting a subset of predictors (tests), given a larger set of possible predictors, in such a manner that a measured criterion can be predicted as efficiently or more efficiently (due to high shrinkage), in the probability sense, from a weighted composite of the subset of predictors as it can be from the whole set of weighted predictors. In other words, the aim of test selection is to maximize prediction of a given criterion and minimize cost where cost is a factor. To accomplish this aim, two common approaches have been followed. These approaches are statistical inference of predictor weights based upon a single sample (Wherry, 1951) and the technique of cross-validation (Mosier, 1951). The cross-validation approach will be studied here because it is the more realistic in the practical sense and because it overlaps with the statistical inference approach.

Implicit in the problem of test selection is the question of differential weighting of predictors, because no test selection can occur without differential weighting. Optimal weighting effectiveness will usually occur when there is considerable variation among the weights of a chosen weighting system (test selection method), when predictor intercorrelations are all small, and when correlations of the chosen weights with corresponding weights of other weighting

systems are small (Scates and Noffsinger, 1931). It can be shown, however, that, under certain conditions, if the number of predictors is very large, the weighted composite will predict a given criterion equally well regardless of how the predictor weights are chosen (Wilks, 1938). Further, the presence of suppressor variables in a battery of predictors may cause ordinary test selection methods to break down during the selection process itself (Wherry, 1946). Thus, it would seem that, in addition to the obvious variable of sample size, the variables of number of potential predictors and homogeneity of potential predictors (as measured by size of test intercorrelations) would be the more important variables to consider in an investigation of the relative efficiency of test selection methods in cross-validation.

The most comprehensive investigation to date of the relative efficiency of test selection methods in cross-validation was conducted by Perloff (1951). Perloff manipulated sample size in his work while holding constant the variables of number of tests in the battery (seven) and homogeneity of the tests (high). His results indicate that the Standard Score Unit Weighting method works best for small samples, Regression Analysis is best for very large samples, and the Wherry Test Selection method works best for medium to large samples. Other less comprehensive comparative studies of test selection methods have been done by Anderson and Fruchter (1960), Grimsley (1949), Lawshe and Patinka (1958), and Lubin and Summerfield (1951). The gist of previous research indicates that the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus one or zero weights only, and the Wherry-Gaylord Test Selection

method allowing plus integer or zero weights only would be the best four methods to study under the manipulation of the variables of sample size, size of test battery, and homogeneity of test battery. Original sources for the Wherry Test Selection method and the two forms of the Wherry-Gaylord Test Selection method can be found in Stead, Shartle, et al. (1940) and Wherry and Gaylord (1946), respectively.

In general, previous studies on comparisons of test selection methods have been based on real data rather than computer generated data. Using computer generated data appears to be better for the present investigation for two reasons. One reason is that the characteristics of computer generated data can be precisely controlled, whereas real data cannot usually be controlled so precisely. The second reason for using computer generated data instead of real data is that the size of the present investigation virtually precludes the possibility of using real data.

A fixed factor, four-way Analysis of Variance with no repeated measures and a within cell n of 5 seems to be an appropriate analysis to use in the present investigation. The four factors are Sample Size, Size of Test Battery, Homogeneity of Test Battery, and Method, denoted by N , S , H , and M , respectively. The levels of N are chosen to be 50 and 200. The levels of S are chosen to be 6 and 12 (does not include the criterion). The levels of H are chosen to be Low and High, as measured by test intercorrelations. The levels of M are the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus one or zero weights only, and the

Wherry-Gaylord Test Selection method allowing plus integer or zero weights only, denoted by WT, RA, WG+1, and WG+I, respectively. The observed data to be analyzed will be cross-validation composite correlation coefficients. The general approach to be used in this experiment is entirely functional in nature. Therefore, no experimental hypotheses will be advanced.

CHAPTER II

METHOD

Subjects. The subjects used in this experiment are entirely fictitious. The computer generated data used are equivalent to 40,000 subjects.

Apparatus. The only apparatus used in this experiment was the computer facilities of The Ohio State University (OSU) Computer Center, consisting of an IBM 1620 Computer and an IBM 7094 Computer.

Procedure. The basic data for the present study were randomly generated intercorrelation matrices having built in N, S, and H characteristics. These characteristics were built in through the use of the 1620 Score Generator Program available at OSU. This program makes use of a theorem in Factor Analysis which says that $Z = AF$ (Harman, 1960), where Z is the matrix of observed scores on each test, A is the population factor structure composed of factor loadings for each test, and F is the matrix of scores on each factor variable, with all factors being assumed to be mutually independent of each other. The program randomly draws samples of pre-assigned size N of factor scores from normally distributed independent populations, and then performs the matrix multiplication A times F to produce the matrix Z of observed test scores for each test. For data generation purposes the criterion variable was considered to be the last test in the matrix

2. The tests could then be intercorrelated to produce the observed intercorrelation matrix of predictors and criterion.

Since the matrix A entirely determines the characteristics of S and H, and since S and H each have two levels, four population matrices A were required for data generation. The number of rows of A controlled S, with 7 and 13 being used (the last row was reserved for the criterion). The pattern of loadings of A controlled H. The patterns of A were so chosen that for the Low level of H the true population intercorrelations of predictors ranged from .010 to .210 with criterion correlations ranging from .050 to .350, whereas for the High level of H a predictor intercorrelation range of .490 to .860 along with a criterion correlation range of .240 to .470 was imposed. The factor loadings for the criterion ranged from .200 to .500 and remained constant across all four A matrices. Furthermore, the additional restriction was imposed that the true population multiple R^2 between predictors and criterion range from .501 to .507 across the four A matrices or, in other words, that it remain essentially constant. Finally, the number of common factors was held constant at four for each of the four A matrices, with unique factors being included for each test. The four A matrices are presented in Table 1.

The proposed 2x2x2x4 ANOVA design with a within cell n of 5 implies that a total of 320 appropriate intercorrelation matrices are required. After the necessary 320 matrices were generated, the four test selection methods under consideration were applied, according to the above ANOVA design, to the first 160 matrices, hereinafter called

TABLE 1
THE FOUR FACTOR STRUCTURES (A MATRICES)^a

Low Homogeneity					High Homogeneity				
Test	Factor				Test	Factor			
	1	2	3	4		1	2	3	4
1	.30	.20			1	.80	.10		
2	.30	.60			2	.90	.30		
3	.20		.60		3	.90		.40	
4	.20		.20		4	.70		.10	
5	.40			.30	5	.80			-0-
6	.40			.10	6	.70			.20
C ^b	.30	.40	.50	.20	C ^b	.30	.40	.50	.20

Test	Factor				Test	Factor			
	1	2	3	4		1	2	3	4
1	.20	.10			1	.80	.20		
2	.10	.10			2	.80	.20		
3	.10	.80			3	.80	.30		
4	.20	.10			4	.80	.40		
5	.30		.10		5	.80		.10	
6	.30		.40		6	.80		.40	
7	.20		.30		7	.80		.20	
8	.20		.10		8	.80		.30	
9	.10			.10	9	.80			-0-
10	.30			.30	10	.80			.20
11	.30			.10	11	.80			.40
12	.10			.10	12	.80			.55
C ^b	.30	.40	.50	.20	C ^b	.30	.40	.50	.20

^aThe unique factor coefficients have been omitted. They may be calculated for any test, i , from the following formula:

$$u_i = (1 - \sum_{j=1}^4 a_{ij}^2)^{\frac{1}{2}}$$

^bC refers to the criterion.

validation matrices, of the 320 matrices to obtain test selection weights. This phase may be called the validation phase of the experiment and its sole purpose is to obtain test weights from each of the 160 validation intercorrelation matrices. Standard criteria for stopping selection of tests were used where applicable. The 160 sets of test weights obtained in validation were then applied to the remaining 160 intercorrelation matrices, hereinafter called cross-validation matrices, using the standard formula for a composite correlation coefficient computed between a criterion and a set of predictors, to obtain cross-validation composite correlation coefficients. This phase of the experiment may be called the cross-validation phase. The data obtained in this phase will form the basis for the ensuing data analysis.¹

¹Since this experiment may be replicated by using the four A matrices together with a score generator program, the generated score data are not available for distribution.

CHAPTER III

RESULTS

It is well known that the technique of ANOVA is relatively robust to a violation of normality of data alone or to a violation of equal cell variances alone (Winer, 1962), but that this robustness may be reduced if both violations are present. The conditional distribution of the random variable $R_c|B$, where $R_c|B$ is the cross-validation composite correlation coefficient given B the observed validation test weight vector, has no simple form except in the case when the true population composite correlation coefficient is zero (Cramér, 1946). It is also known that the distribution of $R_c|B$ departs considerably from normality for moderate sample sizes and that the variance of $R_c|B$ changes radically with sample size (Cramér, 1946). Thus the assumptions of normality and equal cell variances are both violated by the composite correlation coefficient data.

Since $R_c|B$ is in fact a Pearson correlation coefficient, the standard Fisher Transformation, $Z_F = \frac{1}{2} \ln [(1+r)/(1-r)]$, is applicable here. It is known that Z_F is very nearly normally distributed even for rather small samples (Cramér, 1946). The mean of Z_F is very nearly independent of N , the sample size, but the variance of Z_F is equal to $1/(N-3)$ (Cramér, 1946). Since the two sample sizes used in the present experiment are 50 and 200, the ratio of the largest cell variance to the smallest cell variance is about four to one. Such a departure

from equality does not, in the practical sense, affect the ANOVA (McNemar, 1949). Therefore, the Fisher Transformation, $Z_F|B = \frac{1}{2} \ln \frac{(1+R_c|B)}{(1-R_c|B)}$, was applied to the composite correlation data to produce data suitable for analysis which transformed data is presented in Table 2 with decimal points omitted (decimal points appeared in front of each three digit number). A standard fixed factor four way ANOVA was then performed on the data presented in Table 2. The results of the ANOVA are presented in Table 3.

In order to allow for any possible biases in the F tests due to unequal cell variances, a significance point of .025 was set, thereby insuring a true significance point not exceeding .050 in value. The only significant sources of variation are the main effect of N and the triple order interaction of NxSxM. The Newman-Keuls procedure for individual comparisons (Winer, 1962) was applied to the NxSxM effect table. The resulting ordered sums for the NxSxM table are presented in Table 4. An underline joining any two sums in Table 4 indicates that the corresponding means (one-tenth of the sums) do not differ significantly at the .025 point. To aid in interpreting the results, a graphical presentation of the ordered means of Table 4 is presented in Figure 1.

TABLE 2
CROSS-VALIDATION DATA AFTER FISHER TRANSFORMATION

H	S	N	Within Cell	M			
				WT	RA	WG+1	WG+I
LO	6	50	1	250	509	385	389
			2	729	233	589	389
			3	237	422	900	333
			4	211	710	203	587
			5	504	460	539	090
			ΣX	1931	2334	2616	1788
			$\bar{X}$	386	467	523	358
	200	200	1	640	564	390	583
			2	624	469	569	533
			3	450	391	556	559
			4	411	585	295	560
			5	478	568	461	439
			ΣX	2603	2577	2271	2674
			$\bar{X}$	521	515	454	535
	12	50	1	328	482	480	534
			2	497	653	102	296
			3	395	458	358	455
			4	217	333	147	503
			5	459	544	-001	506
			ΣX	1896	2470	1086	2294
			$\bar{X}$	379	494	217	459
	200	200	1	555	527	377	413
			2	556	528	552	569
			3	665	451	591	462
			4	493	484	307	477
			5	348	427	496	461
			ΣX	2617	2417	2323	2382
			$\bar{X}$	523	483	465	476

TABLE 2--Continued

H	S	N	Within Cell	M			
				WT	RA	WG+I	WG+I
H I	6	50	1	437	412	562	438
			2	377	665	504	692
			3	167	769	130	452
			4	400	484	509	353
			5	431	463	720	647
			ΣX	1812	2793	2425	2582
			$\bar{X}$	362	559	485	516
	200		1	482	631	509	458
			2	596	531	337	494
			3	598	499	459	466
			4	713	514	519	561
			5	524	417	455	469
			ΣX	2913	2592	2279	2448
			$\bar{X}$	583	518	456	490
	12	50	1	658	360	357	579
			2	400	400	349	697
			3	219	371	568	417
			4	320	372	525	474
			5	468	193	507	833
			ΣX	2065	1696	2306	3000
			$\bar{X}$	413	339	461	600
	200		1	444	334	469	481
			2	491	538	389	505
			3	383	419	429	409
			4	472	516	528	443
			5	400	497	520	528
			ΣX	2190	2304	2335	2366
			$\bar{X}$	438	461	467	473

TABLE 3
ANOVA SOURCE TABLE

Source	SS	DF	MS	F
N	109,093	1	109,093	7.24 ^a
S	52,237	1	52,237	3.47
H	20,862	1	20,862	1.38
M	61,505	3	20,502	1.36
NxS	1,012	1	1,012	< 1 ^b
NxH	46,596	1	46,596	3.09
NxM	87,373	3	29,124	1.93
SxH	466	1	466	< 1 ^b
SxM	70,352	3	23,451	1.56
HxM	50,589	3	16,863	1.12
NxSxH	5,618	1	5,618	< 1 ^b
NxSxM	146,347	3	48,782	3.24 ^a
NxHxM	56,594	3	18,865	1.25
SxHxM	101,309	3	33,770	2.24
NxSxHxM	102,885	3	34,295	2.28
Error	1,929,174	128	15,072	-
Total	2,842,012	159	-	-

^aSignificant at the .025 point.

^bNot significantly less than one at .025 point.

TABLE 4

Treat.	50	50	50	50	50	200	200	200	200	200	50	200	50	200	50	200
Comb.	(12)	(6)	(12)	(12)	(6)	(6)	(12)	(12)	(12)	(12)	(6)	(6)	(6)	(6)	(12)	(6)
	WG+I	WT	WT	RA	WG+I	WG+I	WG+I	RA	WG+I	WT	WG+I	WG+I	RA	RA	WG+I	WT
Mean	339	374	396	417	437	455	466	472	475	481	504	512	513	517	529	552

^aAny two means jointly underlined are not significantly different at .025 point.

^bDecimals, originally appearing at the beginning of each three digit number, have been omitted.

Fig. 1.--NxSxM Interaction

CHAPTER IV

DISCUSSION AND CONCLUSIONS

There is no need to explain or interpret the significant main effect of N since, according to previous research, this was to be expected. The significant interaction effect of $N \times S \times M$ indicates that sample size, size of test battery, and method jointly interact among themselves. This joint effect can be more clearly grasped by considering what method produces the largest mean at each of the four possible treatment combinations of the levels of N and S. The four possible combinations of N and S are 50(12), 50(6), 200(12), and 200(6), and the best methods for these combinations are WG+I, RA, WT, and WT, respectively. The unexpected result of RA working well for a small sample size and a small number of variables is difficult to explain except on the basis of sampling error. The mean at this treatment combination for WG+I is very little different from that of RA as can be seen in Figure 1. This tends to support the sampling error explanation. The two means are not significantly different from each other according to Table 4.

The order of the four treatment combinations as laid out above appears to be in the same order (descending) of expected overfit, holding method constant. Except for the questionable result for RA, the ordering of best methods corresponding to the treatment combinations as laid out above is about in ascending order of expected

overfit, in regard to method alone. This result is reasonable and indicates that the variable of expected overfit is an important variable in determining what method works best for a given combination of sample size and size of test battery. Another important variable is expected shrinkage which is very closely related to, but not completely determined by, expected overfit (overfit involves only one source of sampling error whereas shrinkage involves two sources of sampling error). It is probable that the operation of these two variables is the chief determiner of what method works best for a given treatment combination. It should be noted that the ranking of methods in regard to expected shrinkage will, in general, be the same as that for expected overfit.

The fact that no significant interaction effects other than $N \times S \times M$ were found in the present study seems to suggest that an extreme sample size alone or an extreme size of test battery alone probably does not produce great differences among methods, or, in other words, that, for this case, any one of the four test selection methods would probably work equally well. Since H did not appear in any significant effect, it must be concluded that the size of predictor incorrelations is probably not an important variable in determining what method to use in a given situation. This is an important practical conclusion since it indicates that it is not necessary to evaluate predictor intercorrelations in determining what test selection method to use in a practical situation.

If the additional variable of cost or effort of method is taken into account, considerable additional information can be

gleaned from the significant $N \times S \times M$ effect. The order (ascending) of the four methods under consideration according to cost or effort is $WG+1$, $WG+I$, WT , and RA , respectively. A consultation of Table 4 shows that the least costly methods, not significantly different from the best methods for the $50(12)$, $50(6)$, $200(12)$, and $200(6)$ treatment combinations, are $WG+I$, $WG+1$, $WG+1$, and $WG+I$, respectively. Thus it would seem that the Wherry-Gaylord Test Selection Method is a far better method than it has been given credit for in the literature. This method also has the distinct advantages that it is equally applicable to test selection or item selection, and that it is easy to understand and apply. Further, the smallest and largest transformed means for the best methods and best least costly methods are .466 and .552, respectively. These values are equivalent to composite correlations of .435 and .502, respectively. In practice, such correlation values would frequently not be considered to differ enough to warrant extra expense.

The interpretation of the significant $N \times S \times M$ interaction has depended heavily upon the assumption that the assumed linear nature of the interaction would still hold if medium-valued levels had been included for N , S , and H . This assumption may not be warranted and should be checked in future research. If H is assumed to appear in no significant effects, the assumption could be investigated by allowing N and S to have three levels, say 6, 9, 12 and 50, 100, 200, respectively, with H being held constant at a medium level. If a within cell n of 5 is used along with the four methods studied in the present investigation, a total of 42,000 fictitious subjects would

be required. Thus the size of this proposed experiment would be about the same as that of the present study.

No negative correlations were allowed in the population intercorrelation matrices used in this investigation. The effects of such correlations should be investigated in future research. If the assumption is made that H would not be involved in any significant effects, than a suitable investigation could be made by modifying the present design such that H is held constant at a medium level and that two additional method levels are added to M. The two new methods would be the Wherry-Gaylord Test Selection method allowing plus or minus unit weights or zero and the Wherry-Gaylord Test Selection method allowing plus or minus integer weights or zero. The population intercorrelation matrices should, of course, have both negative and positive correlations of such a nature that reversing test scoring will not remove the differential sign pattern. If a within cell n of 5 is used, a total of 30,000 fictitious subjects would be required for the proposed design.

CHAPTER V

SUMMARY

The present investigation deals with the relative efficiency of four representative test selection methods in cross-validation, using computer generated data. The four methods were chosen on the basis of previous research and are the Wherry Test Selection method, Regression Analysis, the Wherry-Gaylord Test Selection method allowing plus unit weights or zero only, and the Wherry-Gaylord Test Selection method allowing plus integer weights or zero only, denoted by WT, RA, WG+1, and WG+I, respectively. The efficiency of the four methods was studied under manipulation of the variables of sample size, number of predictors in the test battery, and homogeneity of the test battery (measured by size of predictor intercorrelations), denoted by N, S, and H, respectively, with method being denoted by M. A fixed factor, four-way ANOVA with no repeated measures and a within cell n of 5 was proposed to analyze the data. The levels of N, S, and H were 50 and 200, 6 and 12, and Low and High, respectively. The levels of M were the four chosen test selection methods.

The four methods were applied in validation to randomly drawn intercorrelation matrices with built in appropriate N, S, and H characteristics. The resulting test selection weights were then applied in cross-validation to randomly drawn appropriate intercorrelation matrices to obtain cross-validation composite correlation

coefficients. These coefficients were transformed, by using the standard Fisher Transformation, to insure normality of data. The proposed $2 \times 2 \times 2 \times 4$ ANOVA was then applied to the transformed data, with a significance point of .025 being set in advance. The only significant effects found were the N main effect, which was expected, and the $N \times S \times M$ interaction.

The best single method for the four N, S treatment combinations of 50(12), 50(6), 200(12), and 200(6) were WG+I, RA, WT, and WT, respectively. The result for 50(6) is questionable on a sampling error basis, and it was noted that for this particular treatment combination, WG+I was only slightly less efficient than RA. It was concluded that the joint operation of the effects of expected overfit and expected shrinkage probably determines what test selection method works best for a given treatment combination. The fact that H appeared in no significant effect indicated that size of test intercorrelations is not an important variable to take into account in making a decision as to what test selection method to use in a given practical situation.

If the additional variable of cost or effort of method is taken into account, the best method for the 50(12), 50(6), 200(12), and 200(6) treatment combinations of N and S were WG+I, WG+I, WG+I, and WG+I, respectively. It was concluded that the Wherry-Gaylord Test Selection method, with its advantages of being equally applicable to test or item selection and of being easy to understand and apply, is a far better test selection method than it has been given credit for in the literature.

Since the interpretation of the significant $N \times S \times M$ interaction depended heavily upon the assumption that its assumed linear nature would not change if medium-valued levels had been included for N and S , an experiment was proposed to investigate the validity of the assumption. Further, since no negative correlations were allowed in the population intercorrelation matrices from which sample matrices were drawn, an experiment was proposed to investigate the effects of such correlations upon the efficiency of test selection methods.

REFERENCES

- Anderson, H. E. and Fruchter, B. Some multiple correlation and predictor selection methods. Psychometrika, 1960, 25, 59-76.
- Cramér, H. Mathematical Methods of Statistics. Princeton: University Press, 1946.
- Grimsley, G. A comparative study of the Wherry-Doolittle and a multiple cutting-score method. Psychol. Monogr., 1949, 63(2), 1-24.
- Harman, H. Modern Factor Analysis. Chicago: The University of Chicago Press, 1960.
- Lawshe, D. and Patinka, P. An empirical comparison of two methods of test selection and weighting. J. app. Psychol., 1958, 42(3), 210-212.
- Lubin, A. and Summerfield, A. A square-root method of selecting a minimum set of variables in multiple regression: II. A worked example. Psychometrika, 1951, 16, 425-437.
- McNemar, Q. Psychological Statistics. New York: John Wiley and Sons, Inc., 1949.
- Mosier, C. I. Problems and designs of cross-validation. Educ. and Psychol. Meas., 1951, 11, 5-11.
- Perloff, R. Using trend-fitting predictor weights to improve cross-validation. Unpublished Dissertation, The Ohio State University, 1951.
- Scates, D. E. and Noffsinger, F. R. Factors which determine the effectiveness of weighting. J. educ. Res., 1931, 24, 280-285.
- Stead, W. and Shartle, C., et al. Occupational Counseling Techniques. New York: American Book Company, 1940.
- Wherry, R. J. Test selection and suppressor variables. Psychometrika, 1946, 11, 239-247.
- _____. Comparison of cross-validation with statistical inference of betas and multiple R from a single sample. Educ. and Psychol. Meas., 1951, 11, 23-28.

- Wherry, R. J. and Gaylord, R. H. Test selection with integral gross score weights. Psychometrika, 1946, 11, 173-183.
- Wilks, S. S. Weighting systems for linear functions of correlated variables where there is no dependent variable. Psychometrika, 1938, 3, 23-40.
- Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, Inc., 1962.

APPENDIX

ANNOTATED REFERENCES

Cross-validation

Mosier, C. I. Problems and designs of cross-validation. Educ. and Psychol. Meas., 1951, 11, 5-11.

Mosier distinguishes among five designs closely related to cross-validation as follows:

- "1. Cross-validation - weights determined on one sample and their effectiveness tested on a second, similarly drawn sample.
2. Validity-generalization - weights determined on a sample from one population (and presumably cross validated) but the effectiveness tested against the same criterion on a second sample drawn from a differently defined population.
3. Validity-extension - as in validity generalization except that the criteria as well as the populations differ.
4. Simultaneous validation - item selection or battery weighting on each of two, differently drawn samples in order to develop a single battery useful in either population.
5. Replication - determination of weights in a series of samples drawn from the same population with the final weights (or items selected) based on some combination of the values in the several samples."

Mosier proposes another cross-validation design called double cross-validation in which two samples each serve as a cross-validation sample for the other, thus providing double the usual amount of information.

Differential Weighting and Weighting Problems

Burt, Cyril. The influence of differential weighting. Brit. J. Psychol., Statistical Section, 1950, 3, 105-125.

Burt Discusses the efficacy of differential weighting

under the condition of no external criterion being available and under the condition of an external criterion being available. Burt concludes that when an external criterion is available multiple regression produces the best differential weighting, but, that when no external criterion is available factor analysis, using communalities, yields factors which can then be weighted to produce the best differential weighting. Burt recommends that there is no point in attempting differential weighting if the correlation between the weighted and unweighted predictors is significantly larger than the reliability coefficient for the weighted predictors.

Cureton, E. E. Approximate linear restraints and best predictor weights. Educ. and Psychol. Meas., 1951, 11, 12-15.

It is pointed out that, except in large samples, least-squares estimates of regression weights tend to overfit the data they are based upon, and that the use of these estimates in cross-validation therefore tends to produce composite R's which are less than the multiple R obtained in validation. Cureton further points out that least-squares estimates of regression weights are not necessarily good estimates of population values because of the effects of linear restraints. A linear restraint occurs, for example, when two predictors correlate highly in a sample. Such a high correlation says, in effect, that the two predictors form really only one predictor and the number of predictors is therefore restrained. Such a restraint will cause the estimates of the population weights for the two predictors to fluctuate wildly, with one predictor being weighted highly and the other near zero in a given sample, whereas the reverse may happen in another sample.

Guillikson, H. Theory of Mental Tests. New York: John Wiley & Sons, 1950.

On pages 331-336 of his work, Guillikson discusses the merit or lack of merit of weighting tests directly as their reliability, inversely as their standard deviation, and inversely as their standard error of measurement. Guillikson concludes that it is usually wise to give more weight to a more reliable test but that is difficult to quantify this approach. Guillikson further concludes that weighting tests inversely as their standard deviations should be avoided as a routine procedure, but that weighting tests inversely as their standard error of measurement is probably the safest rule of thumb method to use when no criterion is available and when test intercorrelations have not been computed.

Lev, J. Maximizing test battery prediction when the weights are required to be non-negative. Psychometrika, 1956, 21, 245-252.

An iterative method is developed for computing maximally predictive test weights for a single criterion, under the restriction that only non-negative weights are admissible. Lev shows that a necessary and sufficient condition for the existence of such a set of weights is that the partial derivatives of the maximized multiple R with respect to the weights have values of zero for those tests with positive weights and negative or zero values for those tests having zero weights. Lev's iterative procedure is based upon this theorem. No criterion is given for deciding which of the final set of weights is significant.

Scates, D. E. and Noffsinger, F. R. Factors which determine the effectiveness of weighting. J. educ. Res., 1931, 24, 280-285.

It is concluded that the differential effectiveness of several weighting systems is largely dependent upon three factors. The three factors are variation among themselves of weights in a given system, correlation between corresponding weights of two given systems, and the intercorrelations of predictors. If the variations of weights for each system are all small, then little differential effectiveness of weighting can occur. The same holds true if the correlations among corresponding weights of the systems are all very high or if the intercorrelations of predictors are high. Optimal weighting effectiveness usually occurs when the intercorrelations of predictors are all small, the variation of the weights of the chosen weighting system is high, and the correlations of the chosen weights with corresponding weights of other weighting systems are all small.

Wherry, R. J. Test selection and suppressor variables. Psychometrika, 1946, 11, 239-247.

A suppressor variable in a test battery is a predictor which correlates highly with another, significant predictor, but which itself correlates almost zero with the criterion. Wherry shows how the presence of suppressor variables in a given set of predictors may cause the ordinary test selection methods to break down. The need for suppressor variables in a battery can be eliminated by a prior item analysis of potential predictors against the criterion, to weed out those items having little or no relationship with the criterion. Several other less feasible methods for eliminating the need for suppressor variables are discussed.

Wilks, S. S. Weighting systems for linear functions of correlated variables where there is no dependent variable. Psychometrika, 1938, 3, 23-40.

It is shown that, under certain conditions, the correlation of two linear functions of the same N correlated variables stochastically converges to unity as N approaches

infinity. The larger the mean value of the correlation between pairs of variables, and the smaller the coefficient of variation of the linear coefficients in each linear function, the faster is the convergence to unity of the correlation of the two linear function. Three methods of weighting predictors for a small number of predictors when no criterion is available are presented.

Test Selection Methods and Related Procedures

Garfiel, E. Tests of motor ability. Archives of Psychol., 1923, No. 62, 1-46.

Garfiel discusses in detail on pages 18-21 of his work the Toops Multiple Ratio Method of Test Selection. It is believed that this method is the first test selection method invented by a psychologist; Toop's method involves obtaining approximations to the test weights and multiple R for the best combination of predictors of a given size.

Jenkins, W. An improved short-cut method for Multiple R. Educ. psychol. Measmt., 1952, 12, 316-322.

In this article, Jenkins develops the theory of his test selection weighting method. Several examples are worked out to guide the interested reader. The method will work for any type of non-impossible intercorrelation matrix. Jenkins' method is particularly applicable when one is interested in looking at the multiple R between the criterion and several desired combinations of predictors.

Linhart, H. A criterion for selecting variables in a regression analysis. Psychometrika, 1960, 25, 45-58.

Linhart proposes that the expected length of the 95% confidence interval for the predicted values of the criterion be used as a criterion for deciding whether or not a given subset of predictors in multiple regression should be included in the final test battery. It is shown that the expected length of the 95% confidence interval may decrease if a certain subset of predictors is left out of the analysis, and Linhart provides a statistical test for deciding whether or not to include the certain subset of predictors. The test statistic is a function of the multiple R for the set of all predictors and of the multiple R for the set of predictors with the certain subset left out. The statistic has the same distribution as multiple R with a reduced sample size and a reduced number of independent variables.

Long, W. J. and Burr, I. W. Development of a method for increasing the utility of multiple correlations by considering both testing time and test validity. Psychometrika, 1949, 14, 137-161.

It is pointed out that maximizing the multiple

correlation between a test battery and a criterion on the basis of test validities and intercorrelations alone may produce an impracticable battery because of economic or practical limits on testing time. Long and Burr present a modification of the Wherry Test Selection Method in which tests are selected in the order of greatest rate of return in battery validity per unit of testing time rather than in the order of test contribution to size of multiple R. The modification also makes it possible to determine the largest multiple correlation obtainable from a given battery for a given allowable period of testing time.

Lord, J. M. Efficiency of prediction when a regression equation from one sample is used in a new sample. Research Bulletin (Educ. Testing Service), 1950, RB50-40.

Lord presents a formula for estimating the expected squared errors of prediction in a second sample based upon weights determined in a first sample. The formula is:

$$E_1 E_2 s^2 = [1 + (K+1)/N] \sigma_y^2(x) \text{ where}$$

$$E_1 E_2 s^2 = \text{the expected squared errors,}$$

K = number of independent variables,

N = size of second sample.

$$\sigma_y^2(x) = \text{squared standard error of prediction on first sample.}$$

Ruch, Floyd L. How to Use Employment Tests. Los Angeles: California Test Bureau, 1943.

This reference is the source for the Ruch Multiple Cutting Score Method of test selection. The method involves categorizing the predictor with the highest criterion validity, categorizing the first set of categories on the basis of the variable among the remaining variables, which produces the best differentiation to obtain a second set of categories, categorizing the second set of categories....., etc. The criterion means of the final set of categories are then used as prediction scores for the sample.

Stead, W. and Shartle, C., et al. Occupational Counseling Techniques. New York: American Book Company, 1940.

In the Appendix of this reference will be found the original source (pp. 245-253) for the Wherry Test Selection Method. The method involves the adding of tests to the final battery so long as the value of shrunken R, as computed by the Wherry Formula for Shrunken R, continues to increase significantly.

Summerfield, A. and Lubin, A. A square-root method of selecting a minimum set of variables in multiple regression: I. The method. Psychometrika, 1951, 16, 271-284.

The Summerfield-Lubin Test Selection method is derived. The method is based upon Dwyer's "square root" approach to multiple regression, and differs from the Wherry Test Selection method in that the required computations are more compact, fewer variables are selected, and the multiple correlation and beta weights for the whole battery are computed as a first step with insignificant tests being then thrown out to arrive at the final battery. It is pointed out that the proposed method tends to capitalize on sampling errors.

Toops, H. The L-method. Psychometrika, 1941, 6, 249-266.

This reference is the original source for the Toops' L-method which method is an integral part of the Wherry-Gaylord Test Selection technique. The L-method involves setting up a matrix $[L_{ij}]$ in which L_{ij} is defined to be equal to $\sum X_i X_j - (\sum X_i)(\sum X_j)$, where the summations are carried from 1 to N. From the matrix $[L_{ij}]$ many of the common, statistical calculations are easy to perform.

Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. Ann. math. Statist., 1931, 2, 440-457.

The standard Wherry shrinkage formula for multiple R is derived. The formula is: $\bar{R}^2 = [(N-1)R^2 - (M-1)] / (N-M)$, where $\bar{R}$ is shrunken R, R is the observed value of multiple R, M is the number of independent variables, and N is the sample size.

Wherry, R. J. and Gaylord, R. H. Test selection with integral gross score weights. Psychometrika, 1946, 11, 173-183.

The Wherry-Gaylord Test Selection Method is derived. The method combines Toops' L-method with the Wherry Test Selection Method and results in integral test weights which may be positive, negative, unit or any combination of these. Admission of only integral weights makes the method applicable to either item selection or test selection.

Empirical Comparison of Test Selection Methods and Procedure

Anderson, H. E. and Fruchter, B. Some multiple correlation and predictor selection methods. Psychometrika, 1960, 25, 59-76.

The Doolittle, Wherry, and Summerfield-Lubin methods of test selection are compared both theoretically and in a practical situation. The methods of Wherry and of Summerfield and Lubin are shown to be equivalent. The two methods in a practical situation will select the same set of predictors in the same order provided their "tests" for stopping selection are considered as tools for making decisions rather than

exact statistical tests. Selection by the Doolittle method will usually yield the same set of predictors as the other two methods although in a different order. The Summerfield-Lubin method is recommended over the Doolittle and Wherry methods because of its compactness and ease of computation and also because of the meaningfulness of its interim computational values.

Grimsley, G. A comparative study of the Wherry-Doolittle and a multiple cutting-score method. Psychol. Monogr., 1949, 63(2), 1-24.

The Wherry Test Selection Method and the Ruch Multiple Cutting Score Method are compared for prediction efficiency on seven predictors and one criterion. The two methods were applied to a validation sample of size 250 and were then cross-validated to another sample of size 250. Since the Ruch method produces discrete groups of cases in cross-validation which groups range in mean criterion scores from high downward to low values, comparable groups were obtained for the Wherry method in cross-validation. The means of the comparable groups were compared for the two methods. The differences were all small and none of the corresponding critical ratios were significant at the 5% level. The composite contingency coefficient (cross-validation) for the Wherry method was .455, whereas for the Ruch method the coefficient was .469. The Ruch method required only one-third as much calculation time as the Wherry method, and, at least in the present study, appears to be less affected by shrinkage.

Lawshe, C. and Patinka, P. An empirical comparison of two methods of test selection and weighting. J. app. Psychol., 1958, 42(3), 210-212.

The Jenkins and Wherry Test Selection Methods are compared for the case of 9 predictors and one criterion, using two validation samples of size 184 and 176, respectively, in a typical Personnel Selection situation. Weights were obtained for three tests selected by each method, which were then applied in cross-validation to a sample of size 197. Both methods selected the same three tests, although in a different order, and both methods yielded virtually identical cross-validation composite correlations. The Jenkins method required a great deal less time to apply than the Wherry Test Selection Method.

Lubin, A. and Summerfield, A. A square-root method of selecting a minimum set of variables in multiple regression: II. A worked example. Psychometrika, 1951, 16, 425-437.

A worked example of the Summerfield-Lubin Test Selection method is given, and the method is compared with

the Wherry Test Selection method. The Summerfield-Lubin method appears to be more conservative than the Wherry method in selecting tests.

Perloff, R. Using trend-fitting predictor weights to improve cross-validation. Unpublished Dissertation, The Ohio State University, 1951.

Perloff empirically compared most of the test selection methods which were current in 1951, using a highly homogeneous set of tests. His results indicate that the Standard Score Unit Weighting method works best for small samples, Regression Analysis is best for very large samples, and the Wherry Test Selection method works best for medium to large samples.

Wherry, R. J. Comparison of cross-validation with statistical inference of betas and umltiple R from a single sample. Educ. and Psychol. Meas., 1951, 11, 23-28.

The technique of cross-validation in test selection is compared with the technique of statistical inference of test weights and multiple R from a single sample in test selection. Mathematical statisticians prefer the second technique because definite probability statements can be made. The mathematically naive worker prefers the first technique because one can simply see whether his original results hold up in new samples. Further, the first technique seems to involve no statistical assumptions; the mathematical statistician is, however, unhappy with the lack of probability statements in the first method. Wherry proposes a combination of the two techniques which he hopes might satisfy both parties. The proposed method involves selecting in advance a desired minimum value of shrunken R, computing several composite R's on different cross-validation samples, computing each of the probabilities of a deviation in the desired direction of a composite R from the selected value, and combining, by multiplication, the individual independent probabilities into an overall confidence level for the results.

General References

- Cramér, H. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1946.
- Harman, H. Modern Factor Analysis. Chicago: The University of Chicago Press, 1960.
- McNemar, Q. Psychological Statistics. New York: John Wiley & Sons, Inc., 1949.
- Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, Inc., 1962.

AUTOBIOGRAPHY

I, Lester C. Shine II, was born in Dallas, Texas, April 9, 1933. I graduated from Highland Park Senior High School, Dallas, Texas, in 1950. I received the degree of Bachelor of Business Administration from the University of Texas, Austin, Texas, 1956. After several years in the life insurance business, I entered Southern Methodist University, Dallas, Texas, in 1960. I received the degree of Master of Arts in 1963, with a major in psychology. I then entered the Graduate School of The Ohio State University, Columbus, Ohio, in 1963. Since 1963, I have served as Research Assistant to Dr. J. Naylor, Department of Psychology. At present, I am in the process of completing the requirements for the degree of Doctor of Philosophy at The Ohio State University, with a major in psychology and a minor in mathematical statistics.